Probabilistic opinion models based on subjective sources

Faiza Belbachir, Mohand Boughanem, Malik Missen

Probabilistic Opinion Models Based On Subjective Sources

Faiza Belbachir
University of Toulouse, IRIT
UMR 5505 CNRS
118 route de Narbonne
F-31062 Toulouse cedex 9
Faiza.Belbachir@irit.fr

Mohand Boughanem
University of Toulouse,IRIT
UMR 5505 CNRS
118 route de Narbonne
F-31062 Toulouse cedex 9
Mohand.Boughanem@irit.fr

Malik M.Saad Missen
Department of Computer
Science and IT The Islamia
university of Bahawapur
Pakistan
Saad.missen@gmail.com

ABSTRACT

This article describes approaches for searching opinionated documents for a given query from a standard data collection. To detect if a text is opinionated (i.e., contain subjective information) or not, we propose two methods: the first method is based on lexicons of subjective words (i.e., SentiWordNet) supported by the assumption that more a document contains the subjective terms more it has the tendency of being an opinionated document while the second method is based on probabilistic model supporting the idea that given a document having a strong similarity with a reference opinionated text is more likely to be opinionated. In the second method, we take support of language modeling approach to compute this similarity. Experiments are conducted with TREC Blog06 as the test collection and the IMDB data collection as being the reference data collection. The experimental results report effectiveness of both methods.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Language model, Search process.*

General Terms

Experimentation

Keywords

Information Retrieval, Opinion detection, Blog, Language Model.

1. INTRODUCTION

Opinion mining is the field that deals with extraction of opinions from textual data. It is very different from traditional topic based retrieval where the document returned should not only be relevant to the given topic but must contain opinions about it. The task of opinion detection mining is more difficult and challenging when it is performed on a very casual written text like blogs. In this paper, we propose an opinion detection approach based on language modeling technique and is explained in next section.

2. PROPOSED APPROACH

In this work, for topic-relevance retrieval, we use Okapi BM25 [5] for retrieving top relevant documents for the given topic. For opinion finding we propose two approaches. The first one is based on the assessment of how many terms of a given document are present in a subjective lexicons ($score(d)\_op\_lex$) (see equations 1, 2 and 3) and while the second approach ($score(d)\_op\_pb$) (see equation 9) is based on the direct comparison and computation of a distance between a document and reference collection.

$$score(d)\_op\_lex = \sum_{w \in d} Opinion(w) * P(w|D) \quad (1)$$

$$Opinion(w) = P(w|R) * sub(w) \quad (2)$$

$$Sub(w) = \sum_{si \in sens(w)} \frac{(Neg(si) + Pos(si))}{|sens(w)|} \quad (3)$$

Where $P(w|D)$ and $P(w|R)$ are relative frequency of term w in respectively: document and reference collection. $Neg(si)$, $Pos(si)$ are scores of respectively: Negative and positive of term w as found in SentiWordNet [1].
For the probabilistic based approach we use different language models to represent test document (θ_D) and reference collection (θ_R): Maximum likelihood Models (see equations 4 and 5) and Jelinek Mercer smoothing (see equations 6 and 7).

$$\theta_D = P_{ML}(w|D) = \frac{\#c(w, D)}{|D|} \quad (4)$$

$$\theta_R = P_{ML}(w|R) = \frac{\#c(w, R)}{|R|} \quad (5)$$

$$\theta_D = P(w|D) = \lambda * P_{ML}(w|D) + (1 - \lambda) * P_{ML}(w|R) \quad (6)$$

$$\theta_R = P(w|R) = \lambda * P_{ML}(w|R) + (1 - \lambda) * P_{ML}(w|A) \quad (7)$$

Where $\#c(w, D)$ and $\#c(w, R)$ are frequency of term w in respectively: document and reference collection. $|D|$, $|R|$ designate respectively the length of document D and reference collection R. And $P(w|A)$ is frequency of term w in

analysis collection A. We use the relative entropy measure of Kullback-Leibler divergence [2] to measure the divergence between two probability distributions over the same event space that can be computed as follows:

$$KL - divergence(\theta_D||\theta_R) = \sum_{w \in d} \theta_D * log \frac{\theta_D}{\theta_R} \quad (8)$$

Where θ_D and θ_R are language models respectively: of test document and reference collection (opinion). The opinion score is represent by following equation.

$$score(d)\_op\_pb = \frac{1}{KL - divergence(\theta_D||\theta_R)} \quad (9)$$

More higher is the score, more the document is considered as opinionated. Finally we combine both relevant ($score(d)\_rel$) and opinion score ($score(d)\_op\_lex$ or $score(d)\_op\_pb$) differently by (equation 10 and 11) to obtain final opinion score of the document to the query ($Score(d)_*$ or $Score(d)_+$).

$$Score(d)_+ = \alpha * score(d)\_rel + \beta * score(d)\_op \quad (10)$$

$$Score(d)_* = score(d)\_rel * score(d)\_op \quad (11)$$

Where α and β are combination parameters while $score(d)\_rel$ and $score(d)\_op$ are respectively topic-relevance score and opinion score.

3. EXPERIMENTS AND RESULTS

Experiments are carried out on TREC Blog Track collection [3] with 50 topics of TREC Blog 2006. We use IMDB[1] data collection as reference collection. The results for the experiment are given in tables 1, 2, 3. We use the MAP and P@10 measures and determine the percentage improvement between lexical method and probabilistic methods.

Table 1: The results of final score (Score(d)) using lexical opinion method over TREC topic 2006

Combination	MAP	P@10
Product	0.1306	0.2125
Linear	0.1665	0.3354

Table 2: The results of final score ($Score(d)$) using probabilistic method (Maximum likelihood model) opinion method over TREC topic 2006. Symbol * indicates percentage of improvement between this method and lexical method

Combination	MAP	P@10
Product	0.1898* (45%)	0.4596* (116%)
Linear	0.1999* (20%)	0.4683 *(39%)

On analyzing the above table we see that:

- Linear ranking method is better for combining an opinion and relevance scores (over 5 % of improvement) than Product method.

[1]http://www.cs.cornell.edu/People/pabo/movie-review-data

Table 3: The results of final score ($Score(d)$) using probabilistic method (Smoothing model (JM)) opinion method over TREC topic 2006. Symbol * indicates percentage of improvement between this method and lexical method

Combination	MAP	P@10
Product	0.2198* (68%)	0.5087* (139%)
Linear	0.2294* (37%)	0.5308* (58%)

- Probabilistic method using (Maximum Likelihood or Jelinek Mercer smoothing) models are better than lexical method over (37% MAP and 58% P@10 of improvement) than the first method.
- Probabilistic model using smoothing model improves (MAP 13% and P@10 14%) better than probabilistic using Maximum Likelihood model.
- Probabilistic model using Jelinek Mercer smoothing improves the best results of TREC 2006 participants using the same collection (i.e,. Blog06) and the same topics (i.e., topic06). More than (17% of MAP and 3% of P@10) of improvement compared to the best result of TREC participant [4].

4. CONCLUSIONS

In this article, we proposed a "Compare and Decide" approach to search for opinion documents. To detect if a document contains an opinion, we developped two approaches. The first based on lexical resource relies on how much terms of a reference dictionary are in document and the second based on probabilistic method compute similarity between document and reference collection using language model. Later on, we combined differently an opinion score with relevance score (Product, linear). An improvement was given by the second method (i.e., probabilistic method based on smoothing models) using linear combination. The improvements are 37% for MAP and 58% for P@10 compared with lexical method.

5. REFERENCES

[1] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC'06*, pages 417–422, Genova, 2006.

[2] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of*, SIGIR '01, pages 111–119, New York, NY, USA, 2001. ACM.

[3] C. Macdonald and I. Ounis. The TREC Blogs06 collection: creating and analysing a blog test collection. Number TR-2006-224, 2006.

[4] I. Ounis, M. de Rijke, C. Macdonald, G. Mishne, and I. Soboroff. Overview of the trec-2006 blog track. In *Text Retrieval Conference*, 2006.

[5] S. E. Robertson, S. Walker, M. Hancock-Beaulieu, A. Gull, and M. Lau. Okapi at TREC. In *Text Retrieval Conference*, pages 21–30, 1992.